様式第１１号

論文内容要旨

Mapping mammalian cell-type-specific transcriptional regulatory networks using KD-CAGE and ChIP-seq data in the TC-YIK cell line

学位申請者氏名：Marina Lizio
紹介教員：Hideya Kawaji

第１章　Mapping Mammalian Cell-type-specific Transcriptional Regulatory Networks Using KD-CAGE and ChIP-seq Data in the TC-YIK Cell Line

https://doi.org/10.3389/fgene.2015.00331

Marina Lizio: Frontiers in Genetics Vol. 6 NO. 331, 2015/11/18

Introduction

Regulation of gene expression determines cellular identity and functions. Transcription factors are a special class of genes that have the ability to modulate mRNA levels in a cell until it acquires the predetermined phenotype[1]. However, the full sets of specific transcription factors and their targets are yet undetermined for several cell types. Acquiring such knowledge is fundamental to understanding cellular states, and is applicable to regenerative medicine where efforts are made to engineer or to direct differentiation towards a medically relevant cell type[2].

In order to identify the factors and their direct targets, several approaches have been developed, such as predictive (computational) methods based on the presence of a transcription factor binding site (TFBS) in their promoter regions[3], or experimental methods based on perturbations followed by expression level measurements[4]. However, predictive methods don't perform optimally: TFBS sequences are not well defined for the vast majority of transcription factors, factors from the same family often bind very similar sequences, and binding events may be predicted as important even though a transcription factor is not even expressed in the cell. Similarly, experimental approaches can't discriminate direct from indirect effects of the perturbation. Determination of the physical binding sites in the genome is also possible using protocols such as chromatin immunoprecipitation followed by sequencing (ChIP-seq)[5], but these methods do not distinguish functional from non-functional binding either.

Furthermore, regulatory interactions vary between cell types, as there are different combinations of transcription factors expressed and different chromatin configurations in each cell type. Thus, what we ultimately need is a compendium of regulatory networks specific for every cell type, and we need ways to identify which factors are most important to a given cell type. The FANTOM5 project generated nearly comprehensive sets of promoters with corresponding expression profiles across a large collection of human and mouse samples using CAGE[6,7]. In particular, expression enrichment information (an indication of cell type specificity) for all known transcription factors in a given a cell type are provided, aiding the prioritization of key transcription factors to study cell-type-specific transcriptional regulatory networks (TRNs).

To probe regulatory interactions, we devised an integrated approach for dissecting TRNs using siRNA knock down with CAGE (KD-CAGE) and ChIP-seq in the TC-YIK[8] cervical cancer cell line. TC-YIK is stable, viable and expresses insulin and many other pancreatic genes and transcription factors. Given the difficulty in obtaining primary human beta cells for research, our results may be of interest to studying pancreatic transcriptional regulation.

Results

The TC-YIK cell line expresses pancreatic genes and transcription factors

A systematic review of TC-YIK expressed promoters confirmed the presence of chromogranin-A (*CHGA*), gastrin (*GAST*), insulin (*INS*), ghrelin (*GHRL*) and transthyretin (*TTR*), all playing key roles in the pancreas. Compared to known pancreatic cell catalogues[9], TC-YIK transcriptional profile shows expression for 85% of the beta cell specific genes. Mouse orthologous of most enriched transcription factors in TC-YIK were expressed in at least one stage of pancreatic development also profiled in FANTOM5[10] (**Table 1**).

Symbol	**Expr TPM**	**Enrichmen**	**Pancr/endoc**	**Mouse**	**Experiment**
Transcription factors with enriched expression in TC-YIK					
NEUROD1	593	2.77	yes	yes	Si, CA, CS
INSM1	519	2.72	yes	yes	-
PAX6	296	2.47	yes	yes	Si, CA, CS
NKX6-3	239	2.38	yes	no	-
ARX	237	2.38	yes	yes	Si
MLXIPL	218	2.34	yes	yes	Si, CA
RFX6	146	2.17	yes	yes	Si, CA, CS
ONECUT2	151	2.14	yes	yes	Si, CA
PAX4	133	2.13	yes	yes	Si, CA
PDX1	127	2.11	yes	yes	Si
DACH1	269	2.05	yes	yes	Si, CA
ISL1	102	2.01	yes	yes	Si, CA
FEV	94	1.98	yes	no	Si
HOPX	168	1.95	yes	yes	Si, CA
FOXA2	88	1.95	yes	yes	Si
ST18	78	1.90	yes	yes	-
HNF4G	75	1.88	yes	yes	-
PROX1	106	1.84	yes	yes	Si, CA
HNF4A	69	1.84	yes	yes	Si
ELF3	51	1.71	yes	yes	Si
SHOX2	62	1.70	yes	no	Si, CA
NPAS3	55	1.63	no	yes	-
CDX2	41	1.63	yes	yes	-
HOXA10	40	1.61	yes	no	Si
MNX1	38	1.59	yes	yes	Si, CA
ASCL2	34	1.54	no	yes	-
TFAP2A	97	1.53	yes	no	-
IRF8	31	1.51	no	yes	Si
CASZ1	70	1.51	yes	yes	-
SIX3	30	1.49	no	no	Si
C11orf9/MYRF	62	1.49	no	yes	Si
MYT1	26	1.43	yes	yes	Si
HOXB13	26	1.43	yes	no	Si

ASCL1	25	1.42	yes	yes	Si, CA
NR0B2	24	1.41	yes	yes	Si
LMX1A	24	1.40	yes	no	Si, CA, CS
HSF4	27	1.33	no	yes	-
HES6	71	1.32	yes	yes	-
HLF	23	1.31	no	yes	Si
IRF6	23	1.30	no	yes	-
DLX6	19	1.29	no	no	Si
GATA4	18	1.28	yes	yes	Si, CA
Ubiquitous transcription factors expressed in TC-YIK but not enriched					
ATF5	290	0.73	no	yes	Si, CA
HMGB2	243	0.37	no	yes	Si, CA
GTF3A	213	0.36	no	yes	Si, CA
HMGA1	672	0.34	yes	yes	Si, CA
TBP	29	0.15	no	yes	Si, CA
TAF9	80	0.09	no	yes	Si, CA
TCF25	90	-0.10	no	yes	Si, CA
TAF10	75	-0.33	no	yes	Si, CA

Table 1. List of enriched transcription factors. Experiments abreviations: Si=siRNA; CA=CAGE; CS=Chip-seq.

3.2. Enriched transcription factors are required to maintain the TC-YIK TRN

KD-CAGE[11] identified genome-wide promoters that were perturbed after KD of enriched and non-enriched transcription factors. Looking at TC-YIK enriched promoters only (expression > 3-fold than median across all FANTOM5 samples) we observed a down-regulation of 50% of enriched promoters in the KD of *NEUROD1* and up-regulation of the majority of enriched promoters in *ISL1* KD. Using a measure of anti/pro TC-YIK, defined as the fraction of enriched promoters in the down-regulated set divided by the fraction of enriched promoters in the up-regulated set, we could distinguish anti TC-YIK (high ratio) from pro TC-YIK (low ratio) transcription factors. The majority of enriched and the non-enriched factors appear to be pro-TC-YIK, while *ISL1* and *PROX1* act as antagonists to the TC-YIK state. Interestingly, *MNX1* is pro TC-YIK but appears to do so by actively repressing non-enriched promoters (**Figure 1**).

3.3. ChIP-seq identifies genuine binding sites at promoters and at enhancers

We used ChIP-seq data for four of the TC-YIK enriched factors (*NEUROD1, LMX1A, RFX6* and *PAX6*) to identify their genomic binding sites. Motif enrichment analysis confirmed significant enrichment for the relevant known motifs. For RFX6 there is no known motif; however, the motifs of other RFX family members, and in particular *RFX5*, were overrepresented. *De-novo* motif finding on the *RFX6* ChIP-seq data identified a novel motif that is found in 58% of *RFX6* peaks. All factors often bind together to the same sites, and seldom at promoters. Subsequent comparison of the binding sites to a map of open chromatin sites in human islet cells[9] revealed that between 46% and 62% of peaks overlapped these sites, with preference for enhancer sites (**Figure 2**).

Intriguingly, *RFX6* had twice as many peaks overlapping C5 class enhancers than expected (**Figure 2**, right), suggesting that RFX binding may be one of the earliest events upon sites opening.

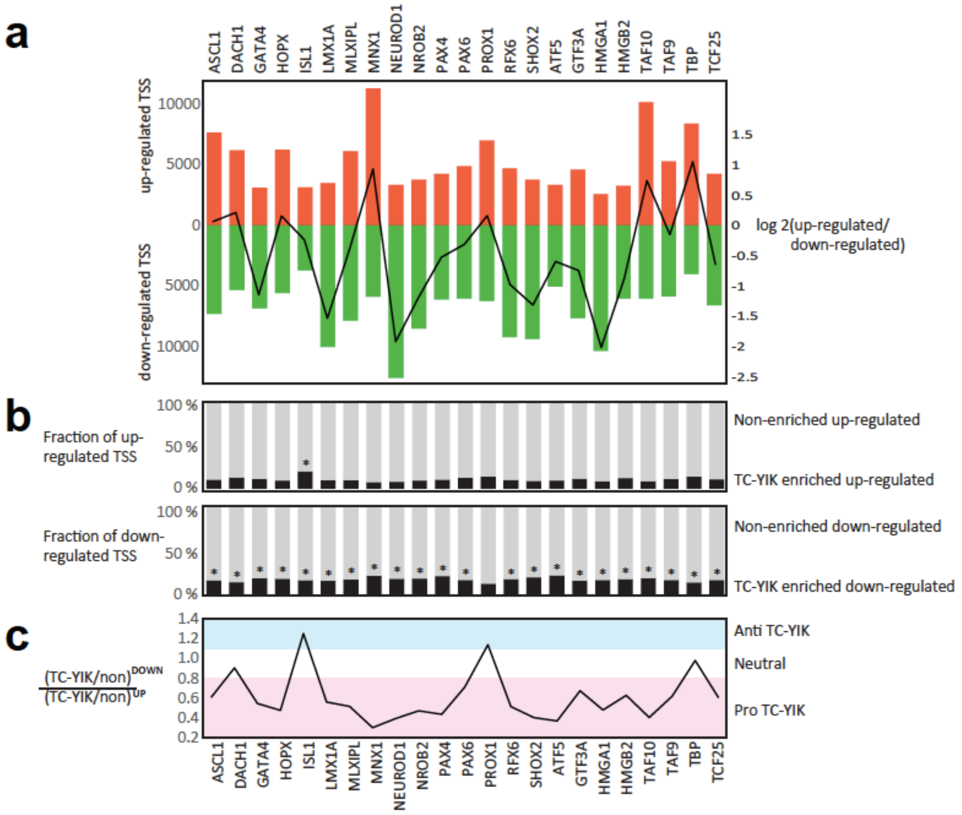

Figure 1. A) Up/down regulated TSSs in KD-CAGE; B) Fractions of up/down-regulated, enriched and non-enriched TSS; C) pro/anti TC-YIK ratios. Adapted from Figure 3 in Lizio et al. 2015.

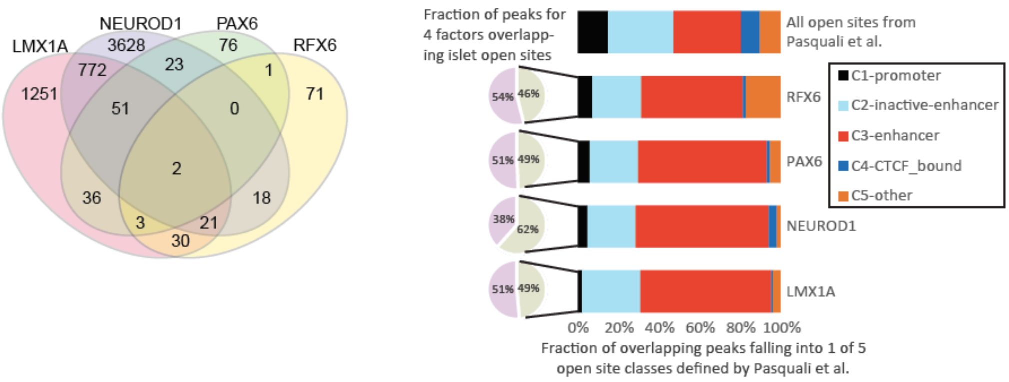

Figure 2. Number of ChIP-seq peaks overlapping across 4 factors (left) and percentages of peaks overlapping open sites (right). Adapted from Figure 4 in Lizio et al. 2015.

Revealing TF-promoters regulatory interaction with ChIP-seq and KD-CAGE

KD-CAGE data combined with ChIP-seq data allowed identification of direct and functional regulation: promoters affected in the KD of a transcription factor that exhibit a ChIP peak of the same factor nearby were considered likely direct targets. We identified 317 and 1,543 directly regulated promoters for *LMX1A* and *NEUROD1* respectively. In particular, NEUROD1 and LMX1A were found targeting directly several other enriched transcription factors in TC-YIK (**Figure3**). Importantly, the promoters within 1kb of a ChIP-seq peak were down-regulated, suggesting that these factors work as activators of their targets. In the case of *RFX6* and *PAX6*, we observed no such distance-dependent effect, suggesting that either these factors work predominantly via distal sites or that the small number of ChIP-seq peaks obtained confounded the analysis. Importantly, not all proximal sites appear to be functional: for *NEUROD1* and *LMX1A* respectively, 17% and 18% of the TSSs within 1kb of a ChIP-seq peak for the same factor were unaffected in the KD. This could mean that such sites are non-functional, or that they are cell-context dependent.

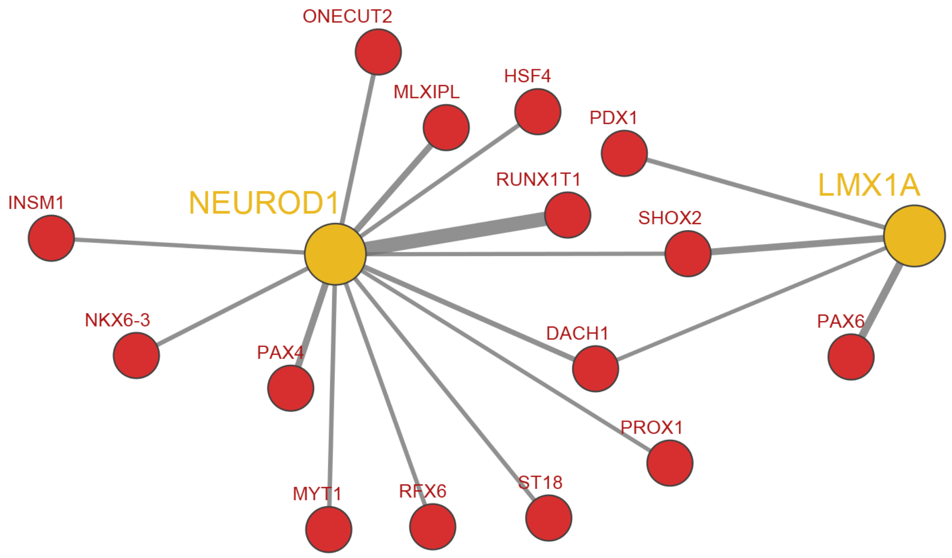

Figure 3. Transcriptional regulatory network of TF-TF direct interactions. Yellow=regulators; red=targets; line thickness=strength of the interaction

Discussion

We presented a method to probe cell type specific transcriptional regulatory networks. We started by identifying cell type enriched transcription factors and then use a combination of siRNA perturbation, CAGE and ChIP-seq to identify their direct and indirect targets that takes advantage of the strengths of all techniques: ChIP-seq can identify bound targets, although it is insufficient to discriminate functional from non-functional bound sites. Conversely, the application of CAGE to perturbed samples identifies affected genes, but cannot distinguish direct from indirect effect. In particular, we stressed on the fact that even in the presence of a protein-DNA interaction, the regulation of target genes can happen only if the site of interaction is functional. This is why complementary techniques should be used in studying TRNs.

Aside from devising a general strategy applicable in several biological scenarios (development, differentiation, or reprogramming) to infer cell type specific TRNs, we have introduced TC-YIK as a model to study transcriptional regulation of neuroendocrine genes.

We have shown that TC-YIK expresses key transcription factors known to be involved in pancreatic cell development and differentiation, that it recapitulates the islet cells transcriptome, and that *NEUROD1, LMX1A, PAX6* and *RFX6* binding sites in TC-YIK are enriched at islet cells active enhancer sites. Thus, such a cell line model could represent a valid vehicle to improve protocols aimed at generating pancreatic beta cells, which are difficult to obtain in terms of quantity, isolation of pure populations, and expansion in culture.

We have shown that not only enriched but also non-enriched factors contribute to the maintenance of the TC-YIK state, as these factors often work cooperatively with state specific factors. The majority of the knock-down experiements revealed a role of these transcription factors as activators; by incorporating ChIP-seq data we could verify their mode of action: for instance, we confirmed that both *NEUROD1* and *LMX1A* work as direct transcriptional activators. In the case of *RFX6* and *PAX6* we made no predictions of their direct targets as there were few peaks bound at promoter regions and there was no enrichment for perturbed TSS near these peaks. This could be due to lower quality or less efficient antibodies used for the two factors, or could reflect lower expression levels compared to the other factors.

Lastly, the study on TC-YIK cell line taught us that mammalian TRN models should incorporate distal regulatory elements as well as proximal elements. That could be achieved by employing chromatin conformation methods to be combined with KD-CAGE and ChIP-seq such that we can identify gold standard regulatory events at both promoters and enhancers, and understand better how each cell type is wired.

References

1. Mitchell PJ, Tjian R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* (1989) **245**(4916):371-8.
2. Cahan P, Li H, Morris SA, Lummertz da Rocha E, Daley GQ, Collins JJ. CellNet: network biology applied to stem cell engineering. *Cell* (2014) **158**(4):903-15. doi: 10.1016/j.cell.2014.07.020.
3. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nature reviews Genetics* (2004) **5**(4):276-87. doi: 10.1038/nrg1315.
4. Consortium F, Suzuki H, Forrest AR, van Nimwegen E, Daub CO, Balwierz PJ, et al. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature genetics* (2009) **41**(5):553-62. doi: 10.1038/ng.375.
5. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods* (2007) **4**(8):651-7. doi: 10.1038/nmeth1068.
6. Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Lassmann T, et al. A promoter-level mammalian expression atlas. *Nature* (2014) **507**(7493):462-70. doi: 10.1038/nature13182.
7. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America* (2003) **100**(26):15776-81. doi: 10.1073/pnas.2136655100.
8. Ichimura H, Yamasaki M, Tamura I, Katsumoto T, Sawada M, Kurimura O, et al. Establishment and characterization of a new cell line TC-YIK originating from argyrophil small cell carcinoma of the uterine cervix integrating HPV16 DNA. *Cancer* (1991) **67**(9):2327-32.
9. Pasquali L, Gaulton KJ, Rodriguez-Segui SA, Mularoni L, Miguel-Escalada I, Akerman I, et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature genetics* (2014) **46**(2):136-43. doi: 10.1038/ng.2870.
10. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome biology* (2015) **16**:22. doi: 10.1186/s13059-014-0560-6.
11. Vitezic M, Lassmann T, Forrest AR, Suzuki M, Tomaru Y, Kawai J, et al. Building promoter aware transcriptional regulatory networks using siRNA perturbation and deepCAGE. *Nucleic acids research* (2010) **38**(22):8141-8. doi: 10.1093/nar/gkq729.